size. Although seemingly more appropriate at first sight, this design is not, in fact, very efficient. It is more reliable to compare the mortality in small villages to the national mortality, and then to discuss the results in light of what is known in terms of geographical variations. Geographical variations in cancer mortality in France are described in detail in Rezvani, Doyon, and Flamant (1986) for the 1971–1978 period and in Rezvani et al. (1997) for the 1986–1993 period.

It is not possible to take into account the age of the reactors since most sites contain several reactors of different generations. We have published our data separately for each site and observed similar results at each. Dr. Bross could certainly salvage some more significant results from these more detailed data (Hill and Laplanche, 1992; Hattchouel, Laplanche, and Hill, 1995).

Our understanding of the way statisticians ought to do science is clearly at variance with Dr. Bross's, and we are not in the least convinced by his undocumented statement that publication of his letter might save thousands of lives worldwide. We shall continue monitoring mortality in the vicinity of French nuclear plants using standard methods.

In the meantime, we maintain the conclusion that our study showed no excess mortality in the population aged 0 to 64 years residing near French nuclear sites.

References

Hattchouel, J. M., Laplanche, A., and Hill, C. (1995). *Mortalité par cancer autour d'installations nucléaires françaises.* Paris: INSERM

Hattchouel, J. M., Laplanche, A., and Hill, C. (1996). Cancer mortality around French nuclear sites. *Annals of Epidemiology* **6,** 126–129.

Hill, C. and Laplanche, A. (1990). Overall mortality and cancer mortality around French nuclear sites. *Nature* **347,** 755–757.

Hill, C. and Laplanche, A. (1992). *La mortalité entre 0 et 24 ans autour d'installations nucléaires françaises.* Paris: INSERM.

Rezvani, A., Doyon, F., and Flamant, R. (1986). *Atlas de la mortalité par cancer en France (1971–1978).* Paris: INSERM.

Rezvani, A., Mollié, A., Doyon, F., and Sancho-Garnier, H. (1997). *Atlas de la mortalité par cancer en France: 1986–1993.* Paris: INSERM, in press.

Catherine Hill and Agnès Laplanche

Truncated Logistic Regression and Residual Intracluster Correlation

From: *Douglas N. Midthune,*
Information Management Services, Inc.
12501 Prosperity Drive, Suite 200
Silver Spring, Maryland 20902, U.S.A.

Edward L. Korn
Biometric Research Branch, National Cancer Institute
Bethesda, Maryland 20892, U.S.A.

Barry I. Graubard
Biometry Branch, National Cancer Institute
Bethesda, Maryland 20892, U.S.A.

To the Editor of Biometrics:

O'Neill and Barry (1995) discuss logistic regression modeling of a binary outcome in the context of a truncated dataset that consists of clusters of observations in which at least one observation in each cluster has a positive response. They compare truncated logistic regression with conditional logistic regression and favor the former because it is more efficient and it allows estimation of cluster-level effects. However, we note that the truncated logistic regression model incorporates the strong model assumption that the logistic regression intercept for each cluster is the same from cluster to cluster; this assumption is unnecessary for the conditional logistic regression modeling. As we will show, there is evidence from the data they analyze that this assumption is violated.

The third column of Table 1 displays the truncated logistic regression coefficients and their maximum-likelihood based standard errors, as given by O'Neill and Barry (1995), for an example involving binary outcomes (dead/alive) for occupants in a car in an accident; only accident data from cars with at least one fatality are recorded in the dataset (111 cars). In addition, we have displayed in the third column of Table 1 the standard errors for the estimated coefficients using robust sandwich-type variance estimators. Such variance estimators have been used for linear regression (White, 1980), generalized linear models (Binder, 1983), and (untruncated) logistic regression analysis of clustered binary data (Zeger and Liang, 1986; Shah, Barnwell, and Bieler, 1996). The robust standard errors in Table 1 are larger than the maximum-likelihood based standard errors. In fact, based on a simulation using the observed pattern of clusters and covariates and the estimated truncated logistic regression coefficients in Table 1, we find that the robust standard errors are significantly larger for four of the variables: Speedlim level 1 ($p = .0022$), Speedlim level 2 ($p = .015$), Speedcat ($p = .018$), and Age level 3 ($p = .020$) (details available from the authors). This type of comparison between maximum-likelihood based and robust standard errors suggests that the model is misspecified (White, 1982).

The fourth column of Table 1 displays the conditional logistic regression coefficients, their maximum-likelihood based standard errors, and their standard errors using robust sandwich-type variance estimators (Fay et al., 1998). The conditional logistic regression does not assume that the intercepts are the same from cluster to cluster. We therefore recommend conditional logistic regression for this example for estimating the association of the outcome with within-cluster level covariates (e.g., sex). The efficiency calculations given by O'Neill and Barry (1995) favoring truncated logistic regression are not relevant here since the data appear to have residual intracluster correlation. We suspect that in most applications there will be questions about the adequacy of a model that assumes no residual intracluster correlation.

What if there is interest in the association of the outcome and cluster-level variables (e.g., damage)? The conditional logistic regression does not help here since the conditioning removes the ability to estimate these associations. One might naively assume that the truncated logistic regression coefficients with their robust standard errors could be used for this purpose; this is the approach that is commonly used with untruncated clustered data analyzed with logistic regression (Zeger and Liang, 1986). However, with truncated clustered data, this approach does not work. This is because the number of clusters with zero events is unknown, so that there is an identifiability problem in estimating outcome/cluster-level associations in the presence of residual intracluster correlation.

Table 1

Truncated logistic regression coefficients, conditional logistic regression coefficients, and random-intercept logistic regression coefficients for single vehicle, frontal impact collisions.[a] *(Maximum-likelihood based standard errors followed by robust standard errors in parenthesis.)*

Variable	Level	Truncated	Conditional[b]	Random intercept
Intercept	1	−3.50 ± .89 (.96)	NA[c]	−9.09 ± 5.35
Resavl	1	1.11 ± .36 (.37)	1.56 ± .46 (.45)	1.58 ± .48
Sex	1	.37 ± .26 (.26)	.41 ± .29 (.27)	.41 ± .31
Age	1	.02 ± .50 (.54)	−.20 ± .55 (.61)	−.11 ± .52
	2	.25 ± .54 (.56)	.31 ± .61 (.66)	.42 ± .57
	3	1.21 ± .72 (.87)	1.38 ± .97 (1.02)	1.69 ± .89
Perloc	1	−.34 ± .31[d] (.31)	−.58 ± .36 (.37)	−.49 ± .36
Damage	1	.53 ± .61 (.60)	NA	.87 ± 1.76
Speedlim	1	.62 ± .52 (.68)	NA	1.22 ± 1.50
	2	.83 ± .83 (1.03)	NA	2.17 ± 2.12
Speedcat	1	1.60 ± .49 (.58)	NA	3.19 ± 1.93
σ		NA	NA	2.20 ± 1.16

[a] Details of the dataset and the variables are as given in O'Neill and Barry (1995).

[b] The estimated conditional logistic regression coefficients differ from those given in O'Neill and Barry (1995) because they artificially reduced the dataset to matched pairs for the conditional logistic regression analysis.

[c] Not applicable.

[d] The difference between the standard error given in O'Neill and Barry (1995) and this one is due to an apparent transcription error in the former.

One approach to this problem is to use a parametric model that explicitly accommodates residual correlation (Barry, 1995). The last column of Table 1 presents the estimated coefficient using a random-intercept logistic regression model in which the intercept is modeled to have a normal distribution with standard deviation σ. Maximum likelihood is used to fit this model to the truncated data. The estimate of σ is 1.9 times its standard error, again suggesting residual intracluster correlation. The estimated coefficients and their standard errors for the individual-level covariates agree fairly well with the conditional logistic regression results. The standard errors of the estimated coefficients of the cluster-level variables are very large, suggesting that there is little information in this dataset concerning these associations.

We wish to thank Dr. Barry for providing his car accident data to us.

References

Barry, C. E. (1995). The regression analysis of group truncated data. Ph.D. dissertation, Australian National University, Canberra.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* **51,** 279–292.

Fay, M. P., Graubard, B. I., Freedman, L. S., and Midthune, D. N. (1998). Conditional logistic regression with sandwich estimators: Application of a meta-analysis. *Biometrics* **54,** 195–208.

O'Neill, T. J. and Barry, S. C. (1995). Truncated logistic regression. *Biometrics* **51,** 533–541.

Shah, B. V., Barnwell, B. G., and Bieler, G. S. (1996). *SUDAAN User's Manual,* Release 7.0. Research Triangle Park, North Carolina: Research Triangle Institute.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test of heteroskedasticity. *Econometrica* **48,** 817–838.

White, H. (1982). Maximum likelihood esimation of misspecified models. *Econometrica* **50,** 1–25.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42,** 121–130.

The authors replied as follows:

Midthune, Korn, and Graubard reiterate many of the cautions and caveats from our paper. We have been very careful to point out the inherent difficulties in modelling truncated data and that very large sample sizes are necessary to give strong conclusions. We clearly state that our analysis should be regarded as exploratory. In the discussion, we imply that, when analyzing data with a strong missing data structure, a variety of techniques can be used to gain insights into the patterns. In our paper, we provide a detailed comparison of both truncated and conditional logistic regression on the road traffic dataset.

Midthune, Korn, and Graubard compare maximum likelihood and robust estimates of standard errors of truncated logistic estimates and find a difference in 4 of the 11 cases in the road traffic application. We have not sighted the details of the calculation of p-values, so we cannot comment on the quoted values.

Barry's dissertation (Barry, 1995) was available to and was cited by Midthune, Korn, and Graubard. In that thesis and in Fay et al. (1998), the inclusion of random effects in the truncated model was considered in detail. The parameter estimates for the random effects model in the final column of Midthune, Korn, and Graubard's Table 1 appear (within rounding errors) both in Barry's dissertation and Fay et al. (1998), which also includes a formal test for random effects. When that test is applied to this dataset, we did not find, after the inclusion of adequate group level covariates, significant random effects (at the .05 level), although there was strong evidence for them.

Our position is that the truncated analysis of this dataset still provides useful information. While we accept that ignoring residual autocorrelation in truncated clustered data such as this will lead to biases, we do not think that these difficulties are unique to this problem. We argue that our analysis still provides more information regarding the group level effects in the data than either performing no analysis or using ordinary logistic regression ignoring the truncation. While biases will exist, their magnitude with respect to the effects is not large enough to totally obscure any relationship. We have simulation results to support this view.